Домашнее задание № 1.

Тема: Оценивание параметров распределения.

Крайний срок сдачи: 7 марта 2023 г., 18:00.

Домашнее задание состоит из четырёх заданий, которые делятся на теоретическую (T1-T4) и вычислительную части (N1-N4). Максимальный балл за домашнюю работу равен 10. В случае обнаружения фактов списывания, всем участникам инцидента ставится оценка -10 (минус 10).

*Решение нужно прислать через бот @aimasters_ bot в виде **одного PDF файла (в любом другом формате решения проверяться не будут)**. Этот PDF файл должен содержать*
- решения теоретических задач T1-T4 набранные в LaTeX, Word,... или написанные от руки и затем отсканированные;
- программный код для численных заданий N1-N4;
- графики, показывающие, что код работает корректно, и численные результаты работы кода.

1

T1 (1.5 балла) Дана выборка из распределения Лапласа (двойного экспоненциального распределения) с плотностью

$$p(x, \theta) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \qquad x \in \mathbb{R}.$$

где $\mu \in \mathbb{R}, \sigma > 0$ - параметры. Найдите оценки параметров μ, σ

(a) методом максимального правдоподобия[1];

(b) методом моментов.

[1]Обратите внимание, что оценка максимального правдоподобия для μ не определяется однозначно в случае, когда размер выборки является чётным числом.

N1 (1 балл) Зафиксируйте значения μ, σ и просимулируйте $M = 100$ выборок размера $n = 999$ из распределения Лапласа. По каждой выборке оцените параметры μ и σ методом моментов и методом максимума правдоподобия. Постройте диаграммы размаха, показывающие, какой метод лучше.

2

Дана выборка $x_1, ..., x_n$ из показательного (экспоненциального) закона с функцией распределения $F(x) = 1 - e^{-\theta x}, x \geq 0$, где $\theta > 0$ - параметр.

T2 (1.5 балла) Рассматриваются оценки параметра θ вида

$$\hat{\theta}_C = \frac{C}{x_1 + ... + x_n},$$

где $C > 0$ может зависеть от n. При каком значении параметра C оценка $\hat{\theta}_C$ является несмещённой? При каком значении параметра C она является состоятельной?

Подсказка. Показательное распределение является частным случаем гамма-распределения $\Gamma(k, \theta)$ с плотностью

$$p(x) = \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x}, \quad x > 0,$$

где $k, \theta > 0$, $\Gamma(k)-$ гамма функция. Известно, что сумма двух независимых случайных величин с распределениями $\Gamma(k_1, \theta)$ и $\Gamma(k_2, \theta)$ имеет распределение $\Gamma(k_1 + k_2, \theta)$.

N2 (1 балл) Известно, что функция надёжности

$$f(t) = e^{-\theta t}, \qquad t > 0,$$

может быть несмещённо оценена при помощи

$$\hat{f}_n(t) = \Big(1 - \frac{t}{x_1 + ... + x_n}\Big)^{n-1} \mathbb{I}\big\{x_1 + ... + x_n > t\big\}.$$

Проверьте это утверждение эмпирически: зафиксируйте θ и смоделируйте выборку размера $n = 1000$ из показательного закона.

Затем для значений t, выбранных по решётке из некоторого интервала, оцените $\hat{f}_n(t)$. Сравните график оценки $\hat{f}_n(t)$ с графиком функции надёжности. Как изменится визуальное впечатление от такого сравнения, если оценить функцию надёжности по нескольким (2-3) выборкам и взять среднее значение полученных оценок в каждой точке?

3

T3 (1 балл) Как было доказано на лекции, для любого набора чисел $X_1, .., X_n$ выборочное среднее $M_X = n^{-1} \sum_{i=1}^n X_i$ и выборочная медиана

$$MED_X = \begin{cases} X_{(k)}, & n = 2k+1, \\ \frac{1}{2}\big(X_{(k)} + X_{(k+1)}\big), & n = 2k, \end{cases}$$

отличаются друг от друга не более чем на (выборочное) стандартное отклонение $\sigma_X = \big(n^{-1} \sum_{i=1}^n (X_i - M_X)^2\big)^{1/2}$, то есть

$$\big|M_X - MED_X\big| \le \sigma_X.$$

Покажите, что данное неравенство не может быть улучшено в том смысле, что не существует константы $C \in (0,1)$, для которой неравенство

$$\big|M_X - MED_X\big| \le C\sigma_X \tag{1}$$

верно для всех наборов чисел $X_1, .., X_n$.

Подсказка. Докажите, что для любой константы $C \in (0,1)$ можно подобрать такое натуральное число m, что неравенство (1) не выполнено для набора из m нулей и $m+1$ единицы.

N3 (1.5 балла) Выборочная медиана является оценкой (теоретической) медианы распределения $x_{1/2}$, которая определяется как решение уравнения $F(x) = 1/2$, где $F-$ функция распределения. Известно, что эта оценка является асимптотически нормальной (см. Лагутин "Наглядная математическая статистика" , глава 7),

$$\sqrt{n}\Big(MED_X - x_{1/2}\Big) \xrightarrow{Law} \mathcal{N}\Big(0, \frac{1}{\big(2p(x_{1/2})\big)^2}\Big), \qquad n \to \infty, \tag{2}$$

при условии, что распределение является абсолютно непрерывным с плотностью $p(x)$ и $p(x_{1/2}) > 0$.

Пусть X имеет показательное распределение со значением параметра $\theta = 5$ (см. задание 2). Перед проведением численного эксперимента, описанного ниже, подсчитайте значение $x_{1/2}$ и значение асимптотической дисперсии $\sigma^2 := \big(2p(x_{1/2})\big)^{-2}$.

(i) Промоделируйте M=100 выборок размера $n = 1000$ из этого распределения.

(ii) Для каждой выборки, оцените левую часть (2).

(iii) Постройте график квантиль-квантиль, сравнивающие эмпирические квантили в левой части (2) с теоретическими квантилями нормального распределения.

(iv) Повторите шаги (i)-(iii) для $n = 10000$, $n = 100000$. Убедитесь, что с увеличением n распределение приближается к нормальному.

(v) Оцените дисперсию выборок при каждом n. Постройте на одном графике три диаграммы размаха дисперсий ошибок (для $n = 1000, 10000, 100000$) и убедитесь, что дисперсии сходятся к асимтотической дисперсии σ^2.

4

T4 (1 балл) Обозначим семейство распределений

$$P_\theta = \Big\{ Law(\xi^2), \qquad \xi \sim \mathcal{N}(0, \theta) \Big\}. \tag{3}$$

(a) Докажите, что данное семейство является экспоненциальным.

(b) Используя только свойства экспоненциальных семейств:

- найдите математическое ожидание и дисперсию величины X, имеющей распределение из семейства P_θ;

- предполагая, что дана выборка $X_1, ..., X_n$ из распределения, входящего в семейство P_θ, найдите оценку параметра θ методом максимального правдоподобия и методом моментов.

N4 (1.5 балла) Пусть $X_0, X_1, X_2, ...$ - цены акций в моменты времени $0, 1, 2, ...$ Рассмотрим квадраты теоретических лог-доходностей

$$Y_k = \big[\log\big(X_k/X_{k-1}\big)\big]^2, \qquad k = 1, 2, ...$$

Рассмотрите цены акции некоторой компании (например, IBM - data(ibm) в пакете waveslim) и для величин Y_k оцените параметры в предположении, что

(i) Y_k имеет распределение квадрата нормальной случайной величины со средним 0 неизвестной дисперсией θ, см. (3);

(ii) Y_k имеет показательное распределение с неизвестным параметром.

Отобразите на одном графике эмпирическую функцию распределения величин Y_k и теоретические функции распределения указанных моделей. Визуально определите, какая из моделей лучше описывает данные.

Подсказка. Обратите внимание, что распределения из семейства (3) являются частными случаями гамма-распределения (см задачу T2), функция распределения которой доступна в языке R (pgamma).